IJAS Iranian Journal of Applied Animal Science

Combined Genome and Protein Statistical Features Improved the Prediction of Genes Encoding Antimicrobial Peptides: A Machine Learning Based Approach

Research Article

M. Rasani[1], K. Karami[1], M. Nassiri[1], M. Tahmoorespur[1] and M.H. Sekhavati[1*]

[1] Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

Received on: 11 Jan 2025
Revised on: 31 May 2025
Accepted on: 11 Jun 2025
Online Published on: Jun 2025

*Correspondence E-mail: sekhavati@um.ac.ir

Online version is available on: www.ijas.ir
https://doi.org/10.71798/ijas.2025.1214792

ABSTRACT

Antimicrobial peptides (AMPs) are increasingly regarded as a promising class of next-generation antibiotics in drug development. Various computational approaches have been developed to predict AMPs, but the majority focus solely on assessing their potency by analyzing the physicochemical characteristics of the peptides. However, to the best of our knowledge, there are no reports that predicted antimicrobial genes in the genome based on characteristics of genomes. In the present study, a novel machine learning-based approach is developed to predict genes encoding AMPs in the genome based on combined physicochemical, genomic, and protein statistical features. Various types of genome features and different machine learning (ML) algorithms are tested to compare the predictive abilities of resulting models. Next, the gene structures of 110 non-AMP and 158 AMP-encoding genes are examined. In this research, 951 genomic and protein features were extracted for AMP and non-AMP-encoding genes in eleven genomic subdomains as well as their 1 kb, 10 kb, and 100 kb upstream and downstream regulatory regions. Among the ML algorithms, the Naive Bayes model processed with an SVM training dataset with an accuracy of 99.63%, precision of 99.41%, recall of 100%, F measure of 99.7%, and area under the curve (AUC) of 1, was identified as the best model. The results showed that due to the heterogeneity of our AMP dataset, using genome features as additional features enhances the performance of all models compared to previous studies that solely relied on AMP sequence-based features.

KEY WORDS antimicrobial peptides, feature selection, machine learning, statistical genome features.

INTRODUCTION

Antimicrobial peptides (AMPs), also referred to as host defense peptides, play a crucial role in the innate immune system of multicellular organisms (Cheema *et al.* 2011; Erdem Büyükkiraz and Kesmen, 2022). These compounds have been used as a model for developing new-generation antimicrobial drugs, considering their diverse structures and broad antimicrobial activities. Antimicrobial peptides are synthesized in mammals by two main pathways: The first pathway is the digestion of larger proteins and the production of short-chain peptides with antimicrobial properties.

The second is the articulation of specific genes encoding antimicrobial peptides at the genome level and, subsequently, ribosomal synthesis of the expressed transcript of these genes (Papagianni, 2003; Schauber and Gallo, 2007; Kondori *et al.* 2011). Some of the peptides in the second category include defensins and cathelicidins. Genes encoding the defensin peptide of neutrophils in humans and rabbits are all about 3 kilobases (kb) pairs in length and comprise 3 exons. The last exon of this group of genes encodes the mature defensin, and the second exon, along with the last exon, encodes the peptide immature form (Martin *et al.* 1995; Selsted and Ouellette, 1995; Ganz, 2007). In this re-

spect, the human and murine enteric defensin consists of two exons, which are similar to the two terminal exons of human neutrophil defensin (Ganz, 2007). The genes encoding the cathelicidin family include four exons: The initial three exons are responsible for encoding the signal peptide. Meanwhile, the last exon contains the sequence encoding the mature peptide with antimicrobial activity, which is created after enzymatic digestion in a specific region (Gudmundsson *et al*. 1995; Zhao *et al*. 1995). The promoter region of this group of peptides contains sites similar to the promoter and regulatory regions of encoding genes in cytokines. These sites indicate that the expression of this group of peptides is likely to be influenced and is in line with the production of cytokines, including Interleukin 6 (IL6) in the incidence of infections (Gudmundsson *et al*. 1995; Zhao *et al*. 1995). In addition to specific encoding genes for antimicrobial peptides in the mammalian genome, there are encoding genes for short-chain peptides that do not have antimicrobial properties (Schlesinger and Elsässer, 2022). This group of peptides may have different metabolic roles, such as hormonal role (Min *et al*. 2012) and immunostimulant (Hancock *et al*. 2016; de la Fuente-Núñez *et al*. 2017). The distinctions between antimicrobial and non-antimicrobial peptides have largely been assessed at the protein level (Bhadra *et al*. 2018; Söylemez *et al*. 2023).

These differences are mainly related to the sequence of amino acids within the peptide structure, physicochemical properties, and specific spatial structures of this group of peptides (Torrent *et al*. 2011). Most of the natural antimicrobial peptides are 10-50 amino acids in length, and their size ranges from 2 to 9 kDa.

Moreover, these positively charged peptides contain highly hydrophobic amino acids (Schauber and Gallo, 2007; Lai and Gallo, 2009). Peptides are classified into specific families according to their amino acid sequences, the presence and identity of certain amino acids, the quantity of cysteine residues, and the arrangement of these cysteines within the sequence (Lay and Anderson, 2005). The three-dimensional structures of antimicrobial peptides (AMPs) are categorized into four main families: α, β, αβ, non-αβ (Wang, 2022). α-helix is the most common structure of AMPs. Disulfide bonds, α-helices, and β strands are types of structural folding. Accordingly, the antimicrobial peptide can interact with the cell membranes of the target pathogens with its proper structure and the greatest ability (Cools *et al*. 2017). Antimicrobial peptides typically consist of positively charged amino acids like arginine (R) and lysine (K), which contribute significantly to their antibacterial effectiveness. This property binds these peptides to negatively charged groups on the surface of bacterial cells, including lipopolysaccharides, which is the first step to destroy the bacterial wall. Antimicrobial peptides, such as defensin, should increase their hydrophobicity to interact with cell bilayer membranes, which are mainly composed of fatty acids and are highly hydrophobic. This improvement is achieved by increasing the proportion of hydrophobic amino acids, such as alanine and cysteine, within their structure (Gasteiger *et al*. 2005; Sagaram *et al*. 2012).

Today, due to the importance of antimicrobial peptides as suitable substitutes for antibiotics, identifying this group of peptides has undergone extensive research. Experimental methods that rely on laboratory techniques for identifying and analyzing new antimicrobial peptides are often expensive and time-consuming. Therefore, computational models, including machine learning techniques, could be very efficient for predicting and providing a faster and more accurate analysis of potential AMP candidates (Khabbaz *et al*. 2021).

The exploration of antimicrobial peptides emerged as a significant area of research in the mid-20th century. This group of activities began with examining secropins of frog and butterfly magainins. Since the 1980s, computational models for quantitative structure–activity relationship models (QSAR models) have been employed to predict and enhance the sequences associated with certain biological activities. Also, since the 1990s, artificial neural networks (ANNs) have replaced conventional QSAR models as efficient machine learning (ML) methods (Papagianni, 2003). Recent progress in machine learning techniques has been utilized to forecast antimicrobial peptides (Khabbaz *et al*. 2021). Based on amino acid physiochemical properties on the protein surface, the predicted toxicity of antimicrobial peptides has been analyzed using linear SVC, random forest, and KNN machine-learning methods. The results are a recall rate of 0.876 and an F1 score of 0.849. The results were evaluated against those generated by machine learning algorithms designed for predicting antimicrobial peptides (AMPs) specific to microbial strains (Vishnepolsky *et al*. 2022). In another study, random forest and AdaBoost had the best performance (Söylemez *et al*. 2023), as ML algorithms were employed for predicting antimicrobial peptides. The outputs are an accuracy of 92%, precision of 92%, recall of 93%, F1 measure of 93%, and AUC of 98%. As mentioned, most of the studies on antimicrobial peptide prediction are based on peptide physicochemical properties (Yeaman and Yount, 2003; Lata *et al*. 2007; Lata *et al*. 2010; Leptihn *et al*. 2010; Vishnepolsky and Pirtskhalava, 2014; Qureshi *et al*. 2015; Cai and Jiang, 2016a; Meher *et al*. 2017; Pane *et al*. 2017). The present study aims to respond to the following questions: Can genomic features be used to identify and predict antimicrobial peptides? To answer this question, we used genomic features in addition to protein-based features to predict genes encoding antimicrobial peptides using artificial intelligence algorithms. For

this purpose, more than 900 known genomic features and motifs at the level of genes encoding antimicrobial peptides, as well as upstream and downstream regulatory regions of these genes, were collected and used as main features along with protein features in the desired algorithms. The results showed that the presence of some genomic features and specific motifs on the DNA level in the genes encoding antimicrobial peptides was more colorful, such that these features can be used as markers for identifying genes encoding antimicrobial peptides. In general, The results of this study indicated that ML methods and the features extracted from the AMPs of three animals (i.e., bovine, sheep, and Gallus gallus) and human species using known genomic features and motifs at the DNA level along with common protein features could significantly improve the accuracy of artificial intelligence-based methods in detecting genes encoding peptides and introduce novel candidate of AMPs.

MATERIALS AND METHODS

The methodology for predicting antimicrobial peptides involves utilizing feature extraction combined with supervised learning techniques, as demonstrated in (Figure 1).

Data collection

The provided content includes a summary of both the positive and negative datasets (Table 1), Positive datasets were peptides whose antimicrobial activity has been experimentally confirmed. The negative dataset was peptides that do not have antimicrobial properties. The dataset used in this study consisted of 158 AMP sequences and 110 non-AMP sequences, which constituted a two-class dataset. The AMP class was assigned a value of 1, while the non-AMP class was designated as 0.

Positive: To construct the positive dataset, we used the genes encoding peptides with antimicrobial properties, whose antimicrobial properties have been confirmed *in vitro* and with a length of 6 to 100 amino acids, from three animal species (i.e., cattle, sheep, and poultry) and humans from the databases. These genes were gathered from publicly accessible databases or datasets. In particular, antibacterial peptides were sourced from the DBAASP database (http://dbaasp.org/home; Jhong *et al*. 2019), CAMP (http://webs.iiited.edu.in/raghava/; Waghu *et al*. 2016), DRAMP (https://dramp.cpu-bioinfor.org; Fan *et al*. 2016), and LAMP (https://ngdc. cncb.ac.cn/database commons/database/id/4562; Zhao *et al*. 2013).

Negative: The negative dataset was constructed by collecting the genes encoding peptides that do not have antimicrobial properties and have a length of 5 to 100 amino acids. The negative dataset for AMP predictors is typically sourced from UniProtKB/Swiss-Prot and databases like the RCSB Protein Data Bank (RCSB PDB), as there is no dedicated database exclusively containing non-AMPs (Liu *et al*. 2017).

Feature extraction

The features used in this study are 268 with 951 characteristics of 158 genes encoding the sequence of antimicrobial peptides and 110 genes encoding the sequence of biological proteins that do not have antimicrobial properties in humans and strategic animals (cattle, sheep, poultry) at the genome level, DNA structure, and transcription level. CpG islands, LTRs, and SINEs, LINEs within the DNA sequence were retrieved from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway; February 2006 build) using its table tools. To identify repeated elements, the UCSC Repeat Masker tracks were utilized. Several attributes, such as the count and frequency of individual amino acids and the distribution of positively and negatively charged amino acids, were analyzed using various bioinformatics tools and software available on the ExPASY platform (http://www.expasy.org). The analysis was conducted across 11 genomic sub-domains: complete genes, exons, introns, 5′-untranslated regions (5′-UTR), 3′-untranslated regions (3′-UTR), as well as sequences located +1 kb, +10 kb, +100 kb upstream and -1 kb, -10 kb, -100 kb downstream. Genomic coordinates were confirmed for the total number of SINEs, LINEs, LTRs, and simple repeats across these regions spanning gene sequence lengths, exons, introns, 5′-UTRs, 3′-UTRs, and the upstream and downstream regions mentioned. These elements were analyzed in relation to genes associated with antimicrobial peptides and those related to various biological proteins. Additionally, the number of CpG islands (CpGi), CpG dinucleotides (CpGn), and the lengths of CpGi regions in these genomic segments were evaluated. Further calculations were performed to determine the density of SINEs, LINEs, LTRs, simple repeats, CpGi, and CpGn per kilobase within the entire gene region sequence, from start to end (Richardson *et al*. 2015).

Using the website https://meme-suite.org/meme/tools/meme, motif sequence, motif length, motif E-value in intron regions, 1kb upstream gene sequence, 1kb downstream gene sequence were collected at the genome level. Next, the percentage of CPG in the sequence of each motif was collected as a feature. The aggregation propensity, both *in vitro* and *in vivo*, was assessed through calculations performed using the AGGRESCAN web server (Table 2) (De Groot *et al*. 2012).

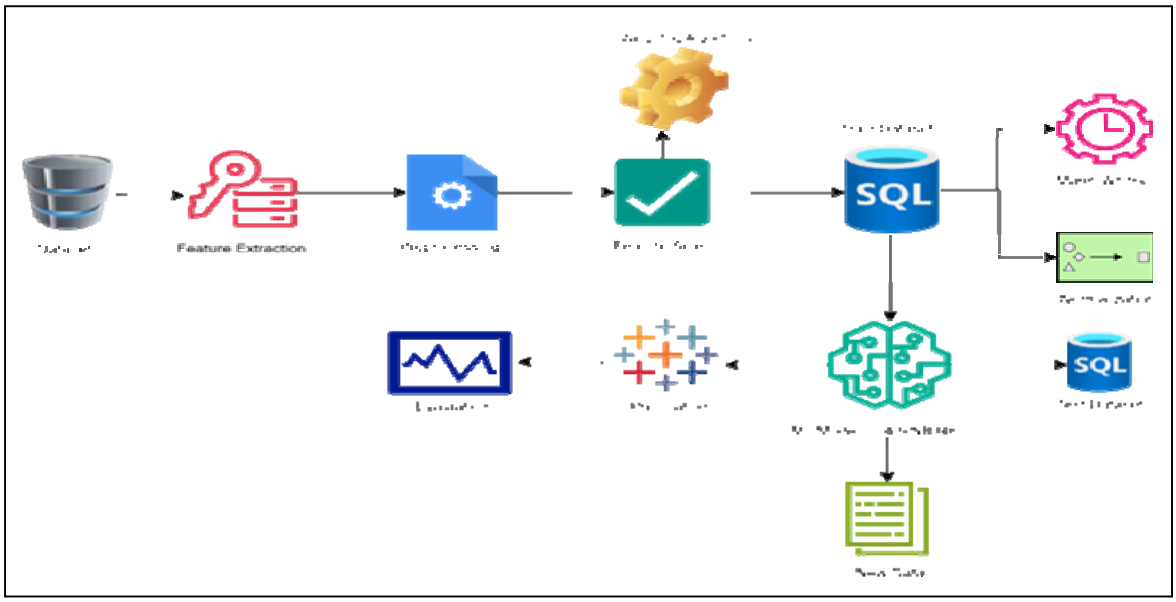

Figure 1 Schematic representation of supervised ML methods in AMP prediction

Table 1 Summary of the positive and negative datasets

Dataset	Antimicrobial, Antibacterial-Antifungal
Positive (AMP)	DBAASP, CAMP, DRAMP, LAMP
Negative (non-AMP)	Non antimicrobial UniProtKB/Swiss-Prot database RCSB Protein Data Bank (RCSB PDB)

Table 2 Categories of features calculated for each peptide

Feature category	Features
Physico-chemical	Net charge
	Charge density
	Isoelectric point
	Normalized hydrophobicity
	Normalized hydrophobic moment
	Hydrophilicity
	Solvation
	Hydropathy
	Amphiphilicity

This research utilized the propypython package to extract a total of 198 features spanning categories such as amino acid composition, pseudo-amino acid composition, and sequence order properties (Cao *et al*. 2013).

Data cleaning

Data cleaning includes finding and righting errors in the dataset, such as dealing with missing or discordant data, removing repetitive, and handling outliers. The data preparation stage includes the following two stages. The first step includes the elimination of data with NA values. In this step, the collected data set has 951 features and 268 genes. Here, we left out 35 features with high noise values (i.e., CpGi 3UTR, length3UTR, CpGi 3UTR, CpGn 3UTR, CpGn 3UTR, Avg CpG 3UTR length, tot CpG 3UTR length, CpGi 5UTR, CpGi exnos/kb, and CpGn introns).

The second step is the elimination of features inappropriate features. In this step, we omitted 143 features because they had inappropriate and irrelevant features. Following the cleaning process, the number of attributes and records was reduced, and the resulting dataset was designated as the finalized clean dataset (CCdb).

Feature selection and attribute weighting

Attribute weighting

The most significant features were determined through the application of five attribute weighting algorithms. The process outlined in reference (41) served as the primary framework for guiding the attribute weighting approach.

Assigning weights based on information gain

This operator determined the contribution of a feature to communication by measuring the information gain in class distribution (Hosseinzadeh *et al*. 2012).

Weight by Gini index

The Gini index algorithm weights each feature through the Gini coefficient. This operator calculates the relevance of a feature by computing the Gini index of the class dispensation if the given example set has been split according to the feature (Hosseinzadeh *et al*. 2012).

Weight by correlation

Algorithm correlation refers to the relationship between a continuous feature and a class feature, evaluated through symmetric uncertainty measurement. When two features exhibit a linear relationship, their correlation coefficient equals ±1. Conversely, a lack of correlation between the features results in a correlation coefficient of 0. Feature weights are determined based on their correlation with the label feature. Features exceeding the predefined threshold value are selected for further processing.

Weight by support vector machine (SVM)

The SVM algorithm is widely used in machine learning, particularly for tasks like classification and regression (Cortes and Vapnik, 1995; Mueller *et al*. 2010).

Weight by optimizing the selection

The genetic weighting algorithm is among the evolutionary-based techniques that are heavily used in feature selection today (Katoch *et al*. 2021).

Feature selection

Feature selection is a process in which the main and important features are selected by discarding the worthless data. Better models can be obtained by using feature selection (Alelyani, 2021) after attribute weighting models that have run on the CCdb. The feature selection stage of the first stage includes adding the data set collected under the name AMP with 773 features and 268 samples to the RapidMiner (RapidMiner 9.10) environment. The second step includes specifying the attribute of the label. We applied feature selection models using feature weighting, including weighting by the information gain, weighting by correlation, weighting by genetic algorithm (selection optimization), weighting by the Gini index, and weighting by SVM. In the Third step, after performing attribute weighting, all variables with weights higher than 0.1 were selected, thereby creating 5 new datasets. The newly created datasets were labeled based on their attribute weighting models (information gain, correlation, Gini index, optimized selection, and SVM) and were utilized in combination with subsequent models, both supervised and unsupervised. After selecting features of high importance, data normalization was performed on these important features. Normalization, as one of the pre-processing processes, is applied to the data set to increase the accuracy of the final model. Therefore, according to the values of the characteristics, the data were normalized in four yields. In the next step, the Mann-Whitney U test (MWU) was performed on highly important features to ensure the appropriateness of important features. Mann-Whitney U (MWU) test was performed between some important features of weighting algorithms. This test belongs to a broader category of tests referred to as non-parametric or distribution-free tests.

Supervised classification

The primary goal of supervised methods is to identify the relationship between input and target attributes. To achieve this, supervised classification was conducted on five newly constructed datasets derived through feature selection. Three classification models C4.5 decision tree, random forest, and Naïve Bayes were applied to these datasets.

Random forest classifier

Random forest is a machine learning model composed of multiple individual decision trees. It functions as an ensemble of tree-based predictors, where each tree is built using a random vector sampled independently but following the same distribution across the forest. Each node signifies a division rule for a specific feature, effectively separating the values in accordance with the chosen parameters (Belgiu and Drăguţ, 2016). Also, it helps identify the genuinely appropriate independent variables so that the system may pick functionality (Breiman, 2001).

Naïve bayes classifier

The Naïve Bayes classifier is a classification method grounded in Bayes' theorem, operating under the assumption that the predictors are independent of one another. It is actually a statistical process based on prediction theory that selects the most likely verdict. Unknown outcomes of identified value systems are estimated by Bayesian probability (Trivedi *et al*. 2022).

Classification trees classifier

Decision trees fall under the category of supervised learning algorithms, with most being constructed through a quantitative minimization approach known as entropy. Among these algorithms, ID3 and C4.5, developed by Quinlan, stand out as the most efficient. C4.5 serves as an advanced iteration of the ID3 algorithm, utilizing an inference method within the decision tree framework. C4.5 is an algorithm invented by Ross Quinlan that is utilized to generate a decision tree on a dataset (Piryonesi and El-Diraby, 2020).

Training machine learning models

A trained classification model is typically evaluated using a hold-out test set to assess its performance on an independent dataset (Figure 2). The accuracy of C4.5, Random Forest, and Naive Bayes models was optimized using the 10-fold cross-validation method. In this method, the training dataset was split into 10 non-overlapping subgroups of approximately equal size.

In each iteration, nine subgroups were used for training, and one subgroup was used for testing. The validation process was conducted a total of ten times. At each validation stage, the performance of models trained in the previous stage was evaluated using the evaluation metrics. The confusion matrix serves as the foundational tool for evaluating the performance of binary classification models, including AMP predictors. This matrix illustrates the four potential outcomes that arise when comparing prediction results to actual class values. These outcomes are categorized and summarized in (Figure 3).

The confusion matrix was derived from the evaluation process, allowing for the calculation of four key parameters: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (Chapelle *et al.* 1999), were calculated by the confusion matrix. In the next stage, seven evaluation metrics of the classification algorithm performance, including specificity, sensitivity, accuracy, precision, recall, F1 score, and AUC-ROC, were calculated using the values obtained from the confusion matrix. At the end of the training and model-building phase, 15 models were built by processing the five training datasets (correlation, optimized selection, information gain, SVM, and Gini index) using three classification algorithms (C4.5, Random Forest, and Naive Bayes).

Sensitivity= TPR= TP / (TP+FN)
Specificity= TN / (TN+FP)
Precision= TP / (TP+FP)
Accuracy= (TP+TN) / (TP+FN+TN+FP)
Classification= (FP+FN) / (TP+TN+FP+FN)

Analysis of the prediction results of the optimal model

After the model-building steps, three methods were used to analyze the prediction results of the built models. The first method involved using an iAMPpred online web server. Considering the broad activity of antimicrobial peptides (e.g., antibacterial, antiviral, antifungal, and anticancer), the trend of 10 genes predicted by the optimal model was checked using the online web service iAMPpred. The second method involved building C4.5, Random Forest, and Naive Bayes models by a known test dataset consisting of 10 genes. The third method involved building C4.5, Random Forest, and Naive Bayes models by a known test dataset consisting of 22 genes. Two subsets were used for validation. In the subset of the first stage, we collected a group of prediction datasets in the human species. This data set of Predict contains 10 compounds (5 antimicrobial peptides whose antimicrobial properties have been confirmed in a laboratory and 5 non-AMP peptides that do not have antimicrobial properties).

The outputs of five weighting algorithms, correlation (correlation), genetic algorithm (selection optimize), information gain (information gain), and SVM (Gini index), were imported as the Train dataset in four species of humans, cattle, sheep, and poultry to the RapidMiner environment. Using three classification algorithms, C45, Random Forest, and Naive Bayes were evaluated and processed on five new datasets created from feature weighting and Predicate datasets. In the subset of the second stage, a group of prediction datasets in the human species was collected. This dataset of Predict contains 22 complexes (5 confirmed antimicrobial peptides and 17 non-AMP that have no antimicrobial properties).

The outputs of five weighting algorithms correlation, genetic algorithm (selection optimize), information gain (information gain), and SVM index (Gini index) were imported as the Train dataset in four species of humans, cattle, sheep, and poultry to the RapidMiner environment. Three classification algorithms, C45, Random Forest, and Naive Bayes, were evaluated and processed on five new datasets created from the feature weighting and Predicate dataset.

RESULTS AND DISCUSSION

The initial data set included 268 AMP sequences and non-AMP sequences, consisting of 951 features after leaving out the features with large empty values (35 cases), of which 143 features were removed due to having inappropriate and irrelevant features. The features used in the present study were reduced to 773 features.

Five datasets were created from each weighting algorithm, which is called the same algorithm. Mann-Whitney U test was performed among the important characteristics in two AMP and Non-AMP groups. The results showed that features such as MOTIF40-1kbUp, MOTIF12-1kbUp, CpGn 100kbUp, Theoretical pI, and Molecular weight that were significant with the Mann-Whitney U test are more important.

The statistically notable distinction between MOTIF42-1kbUp features related to AMP and non-AMP proteins was determined through the Mann-Whitney U test ($P<0.001$). The results of the Mann-Whitney test for each feature, along with the corresponding histograms and box plots, are included in (Figures 4 and 5).

Five new datasets created from feature weighting were processed to train three classification algorithms, C45, Random Forest, and Naive Bayes, to identify genes with antimicrobial properties in four species. Next, these three models were analyzed through the Train dataset. The training was able to predict the Predicate data set.

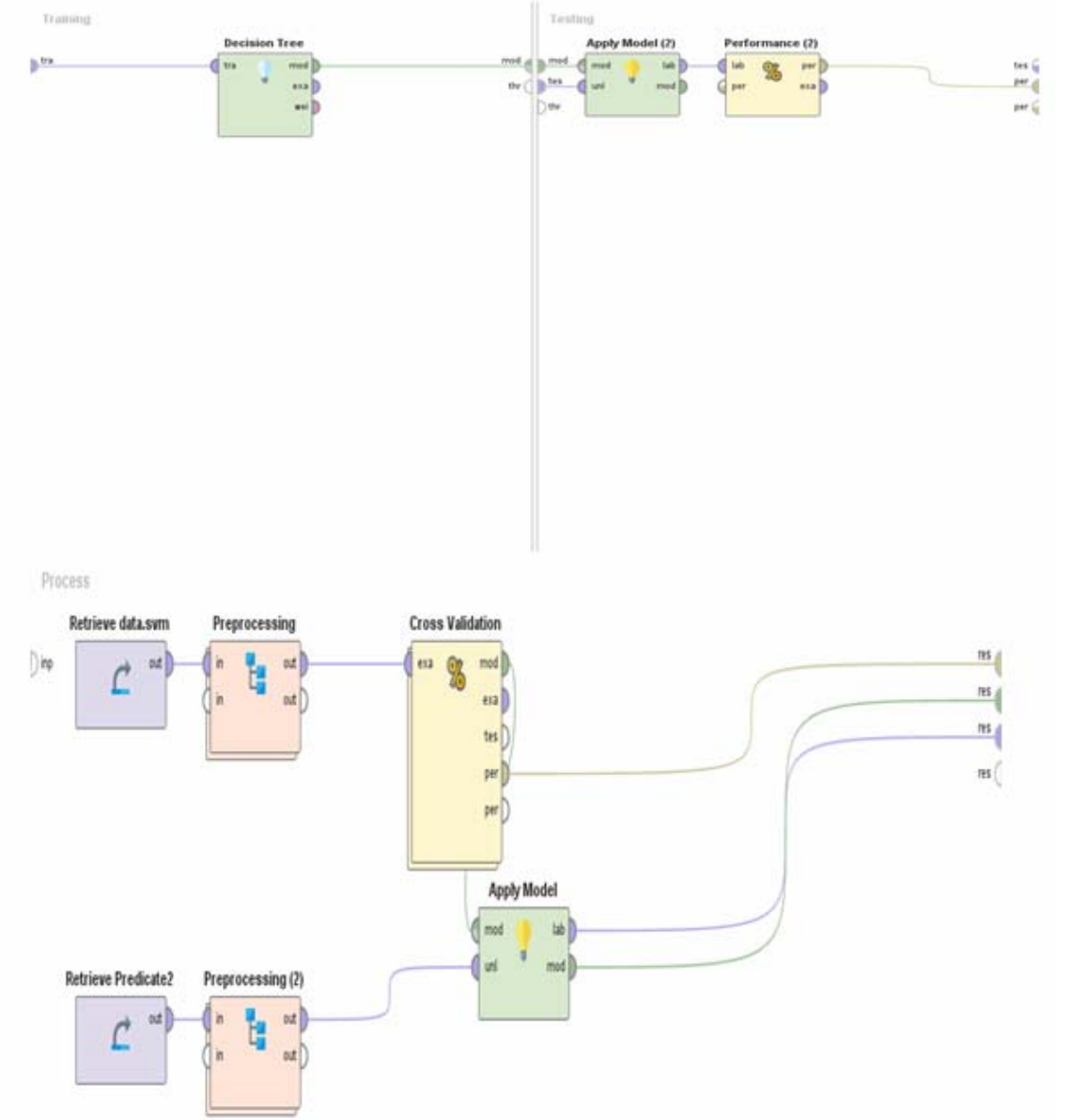

Figure 2 Complete model proposed with RapidMiner

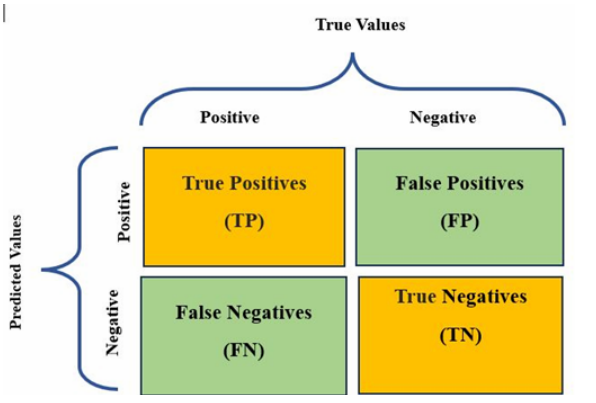

Figure 3 A confusion matrix illustrates the predicted class values generated by a machine learning model in comparison to the actual class values. It effectively highlights the outcomes as true positives, false positives, false negatives, and true negatives

The results of evaluating the output of the weighting algorithm with each of the three classifications were examined (Figure 6). Finally, 15 models were built at the end of the training and model-building stage (i.e., processing of five training data sets correlation, Selection optimization, Information Gain, SVM, and Gini Index using three classification algorithms C45, Random Forest, and Naive Bayes).

Table 3 gives the results of evaluation criteria for building and training 15 Naive Bayes, Random Forest, and C45 classification models by processing five training data sets: information gain, correlation, Gini index, optimized selection, and SVM. Among these 15 developed models, the obtained results are as follows.

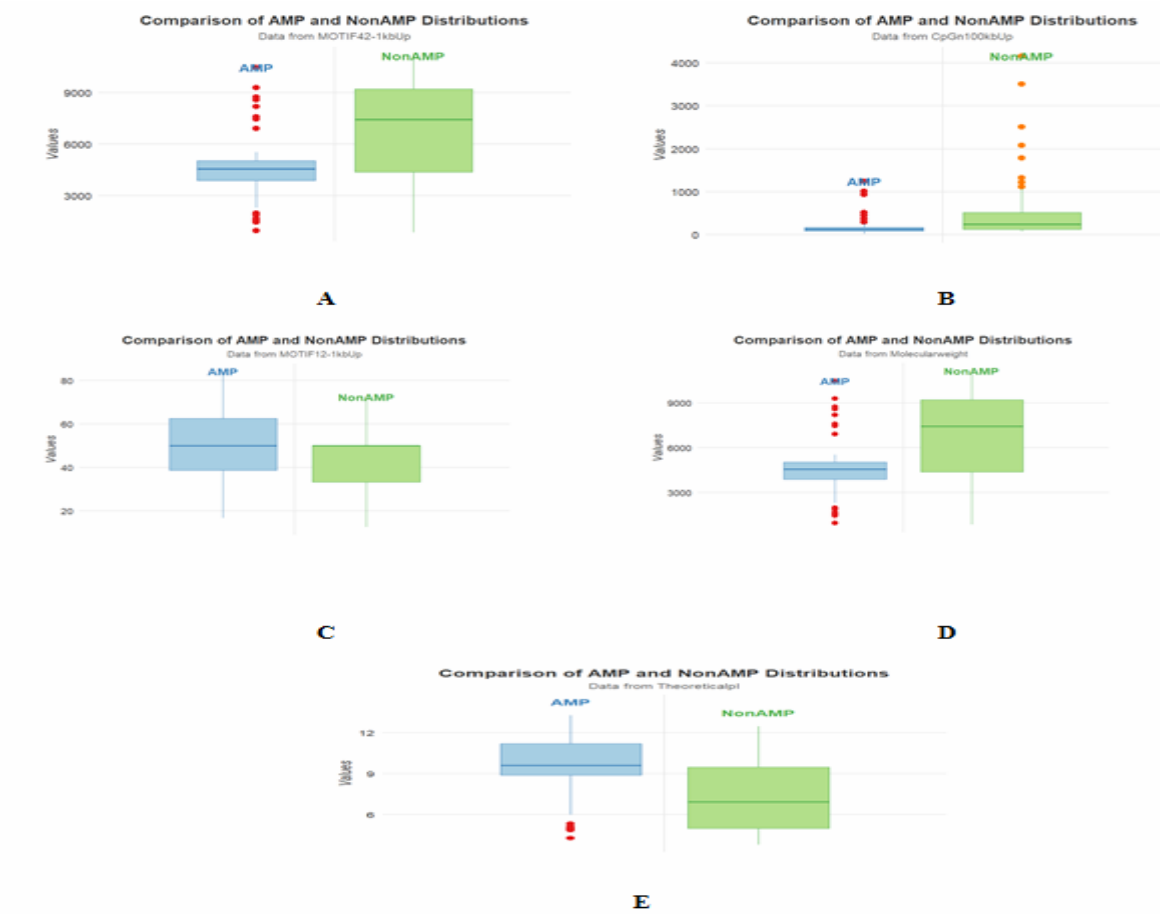

Figure 4 Boxplot and web chart highlighting key attributes of AMPexpressed genes: (a) Boxplot of MOTIF42-1kbUp and P-value (P<0.0001); (b) Boxplot of CpGn100kbUp P-value (P<0.0001); (c) Boxplot of MOTIF12-1kbUp P-value (P<0.0001); (d) Boxplot of Molecular weight P-value (P<0.0001); and (e) Boxplot illustrating the Theoretical PI P-value (P<0.0001) as evaluated using the Mann-Whitney test

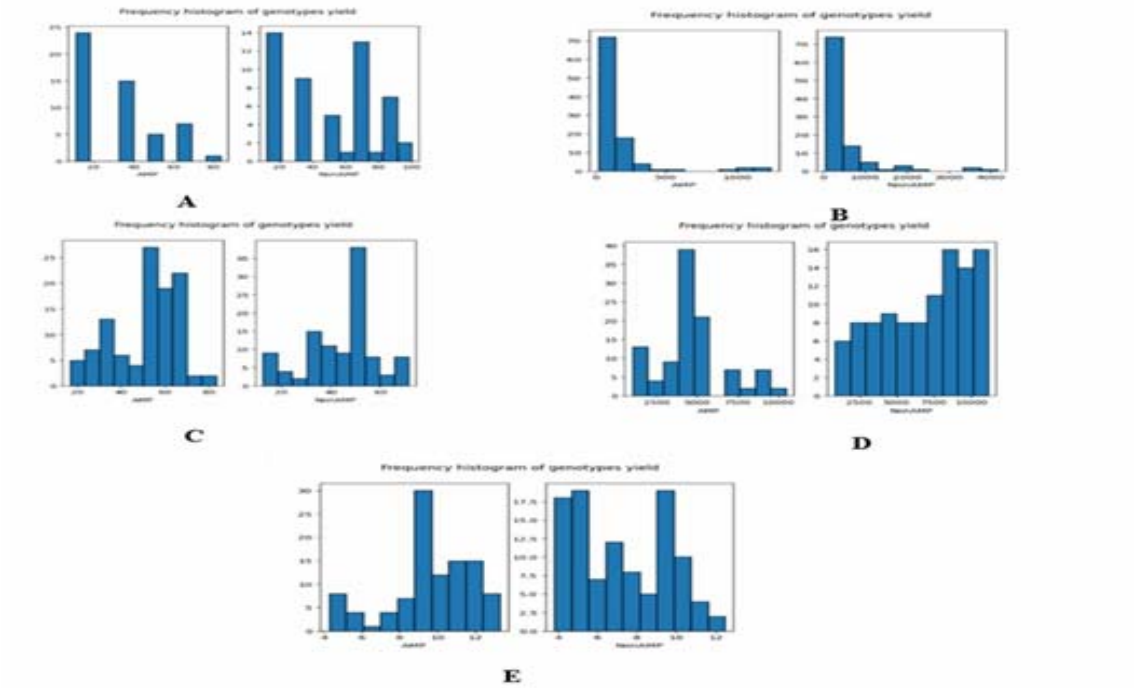

Figure 5 Histogram of some important features of expressed AMP genes. (A) Histogram of MOTIF42-1kbUp, (B) Histogram of CpGn100kbUp, (C) Histogram of MOTIF12-1kbUp, (D) Histogram of Molecular weight (E) Histogram of Theoretical PI

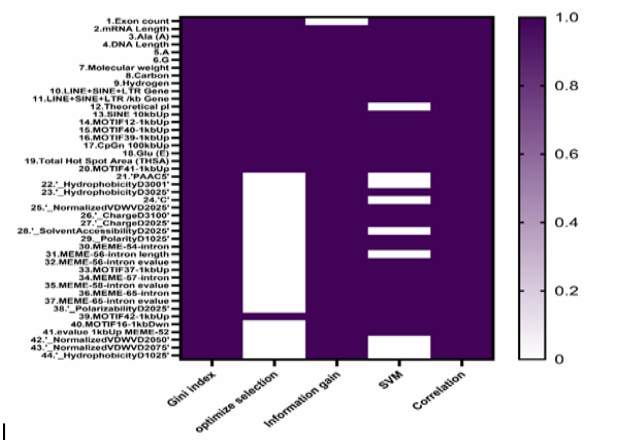

Figure 6 Heat map chart3 of the key attributes identified by various attribute weighting algorithms

In the prediction of AMP coding genes, the results of building and training the Naive Bayes model with SVM weighting algorithms had the best performance with an accuracy rate of 99.63%. On the other hand, building and training the Naive Bayes model with Correlation weighting algorithms had the lowest performance, with an accuracy rate of 94.4% (Table 3).

Since this criterion performs the calculation based on the accuracy of the model in predicting AMP coding genes, the result of building and training the Naive Bayes model, with SVM weighting algorithms, has the best performance with an accuracy rate of 0.37%. Building and training the Naive Bayes model with Correlation weighting algorithms had the lowest performance, with an error rate of 5.23% (Table 3).

In predicting AMP coding genes, the result of building and training the Naive Bayes model with SVM weighting algorithms had the best performance with a 100% completeness rate. Building and training the Naive Bayes model with Information Gain weighting algorithms had the lowest performance with an accuracy rate of 92.38% (Table 3).

In the prediction of AMP coding genes, the evaluation criteria of the F1 score for all the built models was higher than 90%. The result of building and training the Naive Bayes model with SVM weighting algorithms had the best performance, with an F1 score of 99.70% (Table 3).

In predicting the AMP coding genes, the sensitivity evaluation criterion in this research was significant and higher than 91% for all the constructed models. The result of building and training the Naive Bayes model showed the best performance with SVM weighting algorithms with a Sensitivity of 100% (Table 3).

In predicting the AMP coding genes, the sensitivity evaluation criterion in this research was significant and higher than 91% for all the constructed models. The results of building and training the Naive Bayes model with SVM weighting algorithms had the best performance with a Specificity of 99.09%. Building and training the Random Forest model with SVM weighting algorithms had the lowest performance with a Specificity of 93.64% (Table 3). The results of the ROC evaluation criterion from the construction of the C45 model using the SVM algorithm with AUC= 99.9% are given in Figure 7. Also, the results of the ROC evaluation criterion from constructing the Naive Bayes model using the SVM algorithm with AUC=1 are provided in Figure 8. Finally, the results of the criterion ROC evaluation of Random Forest model construction using the SVM algorithm with AUC= 97.6% are shown in Figure 9.

In this research, the iAMPpred online web server was used to predict the specific antimicrobial properties of the obtained 10 genes to confirm the correctness of the ten predicted genes with antimicrobial properties. This online prediction web server is designed to assess the likelihood of peptide sequences functioning as antibacterial, antiviral, or antifungal agents. The output image from the iAMPpred server and the result page for other genes are included in the attached file. Visual snapshots demonstrating the execution process of iAMPpred with a sample dataset, along with corresponding results, are presented in Figure 10. The result page highlights sequences with their probabilities of exhibiting antiviral, antibacterial, or antifungal properties.

The result page displays sequences along with their probabilities of functioning as antiviral, antibacterial, and antifungal peptides. The results obtained from the first stage collection showed five coding genes Amp that were confirmed as Amp antimicrobial properties by each weighting algorithm: correlation (correlation), genetic algorithm (selection optimize), information gain (information gain), and SVM index (Gini index). Also, three classified algorithms, C45, Random Forest, and Naive Bayes, recognized the polypeptides as 100% AMP with 99.63 accuracy (Table 4).

The results obtained from the second stage collection showed five coding genes Amp that were confirmed as AMP antimicrobial properties by each weighting algorithm: correlation (correlation), genetic algorithm (selection optimize), information gain (information gain), and SVM index (Gini index). Here, three classified algorithms, C45, Random Forest, and Naive Bayes, recognized the polypeptide that was 100% AMP with 99.63 accuracy as Amp (Table 5). Today, artificial intelligence (AI) has been increasingly developed in various sciences, including cell biology (Kuiken, 2023). Artificial intelligence is increasingly applied in biological sciences, particularly in predicting the structure and function of genes (Kelley *et al.* 2016; Altman and Krzywinski, 2017; Buchan and Jones, 2019) and proteins (Kelley *et al.* 2016).

Table 3 The Results of evaluation parameters of classification models

	Accuracy	classification errore	AUC	Precision	Recall	F1 measure	FP	FN	TP	TN	Sensitivity	Specificity
Classifier												
C45 Classifier												
Information gain	95.51	4.49	0.954	96.25	96.21	96.19	6	6	152	104	96.21	94.55
Gini index	95.51	4.49	0.954	96.25	96.21	96.19	6	6	152	104	96.21	94.55
Correlation	96.27	3.73	0.961	96.87	96.83	96.83	5	5	153	105	96.83	95.45
Optimize selection	95.11	4.89	0.953	95.75	96.17	95.86	7	6	152	103	96.17	93.64
SVM	97.76	2.24	0.976	97.57	98.75	98.12	4	2	156	105	98.75	96.36
Naive bayes classifier												
Information gain	95.14	4.86	0.976	99.41	92.38	95.69	1	12	146	109	92.38	99.09
Gini index	95.14	4.86	0.976	99.41	100	95.69	1	12	146	109	92.38	99.09
Correlation	94.4	5.6	0.981	99.41	91.13	94.93	1	14	144	109	91.13	99.09
Optimize selection	99.63	0.37	1	99.41	92.38	99.7	1	0	158	109	100	99.09
SVM	**99.63**	**0.37**	**1**	**99.41**	**100**	**99.7**	**1**	**0**	**158**	**109**	**100**	**99.09**
Random forest classifier												
Information gain	98.5	1.5	1	97.61	100	98.77	4	0	158	106	100	96.36
Gini index	99.25	0.75	1	98.79	100	99.37	2	0	158	108	100	98.18
Correlation	98.5	1.5	0.999	97.61	100	98.77	4	0	158	106	100	96.36
Optimize selection	96.64	3.36	0.998	95.83	98.71	97.21	7	2	156	103	98.71	93.64
SVM	97.39	2.61	0.999	95.91	100	97.88	7	0	153	103	100	93.64

Figure 7 The results of the ROC evaluation criteria of the construction of the C45 model using the SVM algorithm

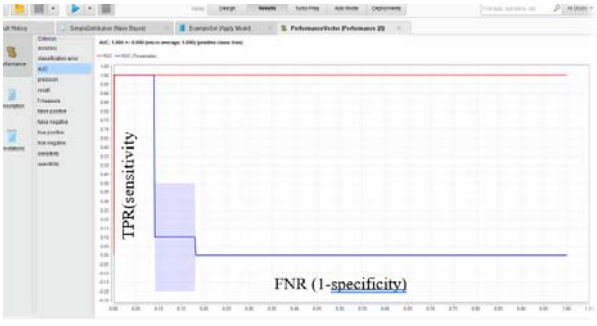

Figure 8 The results of the ROC evaluation criterion of the construction of the Naive Bayes model using the SVM algorithm

Based on the previous research of the present team in the field of antimicrobial peptides, we decided to answer this research question: Can we use artificial intelligence algorithms to predict the genes encoding antimicrobial peptides at the genome level?

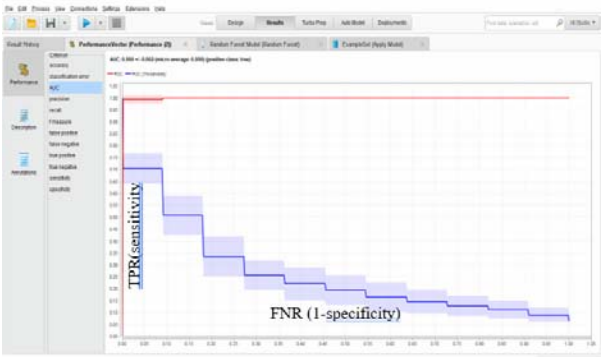

Figure 9 The results of the ROC evaluation criterion of the construction of the Random Forest model using the SVM algorithm

To answer this question in this research, we conducted a comprehensive study of machine learning models to determine and predict genes encoding antimicrobial peptides in humans and three strategic livestock (cattle, sheep, and chicken) using artificial intelligence analysis. In this study, protein and genomic features related to AMP coding genes were simultaneously studied in order to predict AMP coding genes using machine learning algorithms. In previous studies, prediction has been investigated only at the level of amino acid sequences. Understanding the conserved structural features of antimicrobial peptides (AMPs) offers valuable insights into their evolutionary importance and lays the groundwork for developing novel peptide-based antibiotics (Yount and Yeaman, 2004). In this research, a combination of physicochemical features at the protein sequence and DNA sequence levels was used as input in RapidMiner to predict AMPs.

Figure 10 Snapshots of server page of iAMPpred

To this end, features of a large dataset were used. Positive datasets included peptides with experimentally confirmed antimicrobial activity. Negative datasets included short-chain peptides lacking antimicrobial properties. Many previous studies have focused on analyzing a single dataset of antimicrobial peptides (AMPs) sourced from one database (Torrent *et al.* 2011). The AMP sequences utilized in this study appear to be more dependable, as they were sourced from multiple AMP databases. AMPs are crucial for the development and progression of complex multicellular organisms (Zasloff, 2002). The physicochemical properties of this group of active peptides are usually considered acceptable in predicting their performance (Cai and Jiang, 2016b). Physicochemical properties such as hydrophobic sequences (approximately 50%) and net positive charge at physiological pH (Yamagata *et al.* 2003) could indicate AMP biological function (Rončević *et al.* 2019). One of the most important advantages of the work done with the previous work described in the background review section is the feature selection step. At this stage, the selection and weighting of important features from the AMP data set were done separately with 5 feature selection algorithms: gain information, correlation and SVM, Gini index, and selection optimization. We also examined the aggregation tendencies of both groups, AMP and non-AMP, as these could serve as significant modulators of peptide function.

Furthermore, AGGRESCAN serves as an effective tool for analyzing aggregation in bacteria. Similar to previous studies, Torrent *et al.* (2011) conducted a study on linking the physicochemical and antimicrobial properties of peptides through a rational prediction model, utilizing the AGGRESCAN analytical tool (Zasloff, 2002). The present was conducted using a machine learning-based approach, correlating a complete set of physicochemical properties derived from peptide and DNA sequences and genome-level motifs with antimicrobial activity. Although physicochemical properties have been used in previous studies (Meher *et al.* 2017; Bhadra *et al.* 2018), CG percentage was obtained in the sequence of each motif. In a recent study, we were able to predict the genes encoding antimicrobial peptides with very high accuracy (99.63%) by considering the motifs at the genome level as a genomic feature in addition to the physicochemical features of the protein (Table 6). Characterization of genome-wide motifs was used for the first time in this study to predict AMP. Protein motifs are small sections within a protein's three-dimensional structure or amino acid sequence that are commonly found across different proteins. These regions are identifiable portions of the protein structure and may, but do not always, serve a distinct chemical or biological purpose (Richardson, 1994).

Table 4 The performance of fifteen distinct induction models was evaluated based on eight criteria accuracy, classification error, AUC, precision, recall, F-measure, sensitivity, and specificity using 10-fold cross-validation across 10 datasets

	Accuracy	Classification Errore	AUC	Precision	Recall	F meas-ure	FP	FN	TP	TN	Sensitivity	Specificity
C45												
Information gain	95.51	4.49	0.954	96.25	96.21	96.19	6	6	152	104	96.21	94.55
Gini index	95.51	4.49	0.954	96.25	96.21	96.19	6	6	152	104	96.21	94.55
Correlation	96.27	3.73	0.961	96.87	96.83	96.83	5	5	153	105	96.83	95.45
Optimize selection	95.11	4.89	0.953	95.75	96.17	95.86	7	6	152	103	96.17	93.64
SVM	97.76	2.24	0.976	97.57	98.75	98.12	4	2	156	106	98.75	96.36
Naive bayes												
Information gain	95.14	4.86	0.976	99.41	92.38	95.69	1	12	146	109	92.38	99.09
Gini index	95.14	4.86	0.976	99.41	92.38	95.69	1	12	146	109	92.38	99.09
Correlation	94.4	5.6	0.981	99.41	91.13	94.93	1	14	144	109	91.13	99.09
Optimize selection	99.63	0.37	1	99.41	100	99.7	1	0	158	109	100	99.09
SVM	**99.63**	**0.37**	**1**	**99.41**	**100**	**99.7**	**1**	**0**	**158**	**109**	**100**	**99.09**
Random forest												
Information gain	98.5	1.5	1	97.61	100	98.77	4	0	158	106	100	96.36
Gini index	99.25	0.75	1	98.79	100	99.37	2	0	158	108	100	98.18
Correlation	98.5	1.5	0.999	97.61	100	98.77	4	0	158	106	100	96.36
Optimize selection	96.64	3.36	0.998	95.83	98.71	97.21	7	2	156	103	98.71	93.64
SVM	97.39	2.61	0.999	95.91	100	97.88	7	0	158	103	100	93.64

Table 5 The performance of fifteen distinct induction models was evaluated based on eight criteria accuracy, classification error, AUC, precision, recall, F-measure, sensitivity, and specificity using 10-fold cross-validation across 22 datasets

	Accuracy	Classification Errore	AUC	Precision	Recall	F meas-ure	FP	FN	TP	TN	Sensitivity	Specificity
C45												
Information gain	98.87	1.13	1	98.2	100	99.07	3	0	158	108	100	97.27
Gini index	94.77	5.23	0.981	99.41	91.75	95.27	1	13	145	109	91.75	99.09
Correlation	97.38	2.62	0.972	97.53	98.08	97.78	4	3	155	106	98.08	96.36%
Optimize selection	97.38	2.62	0.999	95.81	100	97.84	7	0	158	103	100	93.64
SVM	98.87	1.13	1	98.2	100	99.07	3	0	158	107	100	97.27
Naive bayes												
Information gain	98.87	1.13	1	98.2	100	99.07	3	0	158	107	100	97.27
Gini index	94.77	5.23	0.981	99.41	91.75	95.27	1	13	145	109	91.75	99.09
Correlation	94.77	5.23	0.981	99.41	91.75	95.27	1	13	145	109	91.75	99.09
Optimize selection	99.63	0.37	1	99.41	100	99.07	1	0	158	109	100	99.09
SVM	**99.63**	**0.37**	**1**	**99.41**	**100**	**99.7**	**1**	**0**	**158**	**109**	**100**	**99.09**
Random forest												
Information gain	98.87	1.13	1	98.2	100	99.07	3	0	158	107	100	97.27
Gini index	98.87	1.13	1	98.2	100	99.07	3	0	158	107	100	97.27
Correlation	99.25	0.75	1	98.79	100	99.37	2	0	158	108	100	98.18
Optimize selection	97.38	2.62	0.999	95.81	100	97.84	7	0	158	103	100	93.64
SVM	97.38	2.62	0.999	95.81	100	97.84	7	0	158	103	100	93.64

Table 6 AMP predicted by supervised ML methods

Source	Gene	Source	Gene
Bovine	ATP5MD, DAPIT, USMG5	Bovine	COX7A1, COX7A, COX7AH
Gallus gallus	UQCRFS1, RCJMB04_5b19	Bovine	BGLAP
Homo sapiens	COD, SNRPB1	Bovine	MRPL33
Homo sapiens	UBL5	Bovine	MRPL34
Homo sapiens	POLR2K	Bovine	MRPL27

DNA sequence motifs are short sequences that are present in different parts of DNA. The binding site of transcription factors in DNA sequences is determined using motifs (D'haeseleer, 2006). The most innovative aspect of this method lies in considering genome-level motifs within the intron region, as well as 1 kb upstream and downstream of the gene, as influential features. It takes into account how gene expression is affected by the nucleotide composition of the coding region, including factors such as GC content and codon usage (Barahimipour *et al*. 2015).

Regarding motif features, previous studies have shown that the Gly-Xaa-Cys motif is conserved across all mammalian defensins and plays a crucial role in ensuring the proper folding and natural structure of antimicrobial peptides (AMPs) (Xie *et al*. 2005). Disulfide bridges, found in antimicrobial peptides (AMPs) like formicin, drosomycin, protegrin-1, big defensin, gaegurin-1, polyphemusin-1, mytilin A, gomesin, HNP-3 ,thanatin, and AFP-1, typically include conserved GXC or CXG motifs (Yount and Yeaman, 2004). LL-37 is an amphipathic peptide featuring a helical structure, with its heparin-binding property attributed to the XBBXBX motif (Andersson *et al.* 2004). Antimicrobial peptides (AMPs) capable of binding to heparin often feature specific heparin-binding motifs, such as XBBBXXBX, where X signifies hydrophobic or uncharged amino acids, and B represents critical basic residues. Those containing proline-arginine-proline (PRP) motifs belong to a category of proline/arginine-rich cationic peptides, which includes examples like callinectin and astacidin 2. These peptides commonly exhibit one or more PRP motifs and demonstrate strong antibacterial effects against both Gram-positive and Gram-negative bacteria. Another notable example is armadylidine, a glycine-rich cationic antimicrobial peptide characterized by the presence of five repeated GGGFHR or GGGFHS motifs and C-terminal amidation. This peptide shows significant antibacterial efficacy, particularly against Gram-positive bacteria (Herbinière *et al*. 2005). Histatin peptides are histidine-rich peptides belonging to a family of AMPs. The human salivary peptide, histatin 5, along with other histatins, features an ATCUN motif. The antifungal activity of histatin 5 is notably associated with generating activity of reactive oxygen species of the Asp-Ser-His motif (Cabras *et al*. 2007; Tay *et al*. 2009).

Previously, various papers on bioinformatics have investigated motifs related to the immunological properties of peptides for several purposes (Attique *et al*. 2020). At the genomic level, peptides derived from natural sources usually show common sequence motifs (Cardoso *et al*. 2021). In previous studies, Boris *et al*. (2022) performed a comparative evaluation of machine learning algorithms focused on predicting antimicrobial peptides (AMPs) specific to microbial strains. Incorporating genome characteristics as additional features improved the performance of all models compared to those based solely on AMP sequence properties (Attique *et al*. 2020). Marina *et al*. (2023) examined the γ-Core Motif Peptides of Plant AMPs, highlighting their potential as innovative antimicrobials for both medicine and agriculture. Their findings revealed that the γ-core motif possesses inhibitory effects against bacterial and fungal pathogens affecting plants and humans, making it a promising framework for creating new anti-infective agents. Additionally, certain γ-core peptides demonstrate a dual function by combining antimicrobial activity with immunomodulatory properties, further expanding their range of practical applications (Attique *et al*. 2020).

A recent study revealed, some properties had the highest weight in all outputs of all five weighting algorithms. In our work, these properties were evaluated in antimicrobial and non-microbial peptides to test their suitability for AMP prediction. The Mann-Whitney U test results showed a significant difference in MOTIF42-1kbUp ($P<0.0001$), theoretical pI ($P<0.0001$), MOTIF12-1kbUp ($P<0.0001$), molecular weight ($P<0.0001$), and CpGn100kbUp ($P<0.0001$) (Karami *et al*. 2019). In a study, CpG (CpGn) was used to identify imprinted genes in bovine. Gene expression was influenced by the nucleotide composition of the coding region, such as GC content and codon usage (Karami *et al*. 2019). As LINEs and SINEs were introduced as genomic properties in previous studies by (Cowley *et al*. 2011), SINE and LINE sequences were employed as genomic properties for prediction in the present study (Cowley *et al*. 2011). The SINE sequence contains 13% of the total human genomic DNA (Vassetzky and Kramerov, 2013). These sequences do not code for a protein. Long interspersed nuclear elements (LINEs) represent a class of non-long terminal repeat (non-LTR) retrotransposons that are

widespread across the genomes of many eukaryotic species. Also, the physicochemical property of charge was considered the key factor for antimicrobial activity.

In the study by Bhadra *et al*. (2018) this property was used to predict AMP. Moreover, protein aggregation tendency, which could be an important modifier of peptide function, was investigated in our research. The protein's tendency to aggregate largely depends on the polypeptide composition and primary structure. Some parts of very short amino acid sequences could act as aggregation facilitators (Ivanova *et al*. 2004; Ventura *et al*. 2004), which is consistent with previous studies in using protein aggregation as a property (Torrent *et al*. 2011; Khabbaz *et al*. 2021). The prediction method used in our study provided the best AUC result (AUC=1). For this purpose, the 10-fold cross-validation method was used. AUC is considered among the criteria that could be examined for AMP prediction. The partial AUC is the specific area under the ROC curve (Peterson *et al*. 2008). Several computational techniques have been recently proposed to identify and synthesize antimicrobial drugs as well as accelerate the selection of suitable candidates from biological compounds. DNA sequence-based and protein sequence-based models have been trained using machine learning algorithms as the main techniques to distinguish AMPs from non-AMPs. Thomas *et al*. (2010) suggested using supervised learning methods such as random forest (RF), support vector machines (SVMs) and discriminant analysis (DA) to predict AMPs based on physiochemical properties of amino acid sequences. Their prediction models had the accuracy of 93.2%, 91.5%, and 87.5% for RF, SVM, and DA, respectively (Lata *et al*. 2010). In another study, an SVM-based model was developed by combining the N and C termini of the complete peptide amino acid as one of the physicochemical properties used to predict protein antibacterial peptides through an amino acid sequence. They reported a model accuracy of 92.14% and concluded that antibacterial peptides prefer certain residues at N and C termini, contributing to distinguish them from non-antibacterial peptides (Bhadra *et al*. 2018). Elsewhere, amino acid physicochemical properties were utilized to predict the sequences of antimicrobial peptides. By leveraging the distribution patterns of these properties combined with a random forest algorithm, the approach achieved a prediction accuracy of 96% (Lee *et al*. 2016). In another research, an SVM classifier-based predictive model of AMP was developed using machine learning algorithms. The results showed the prediction accuracy, specificity, and sensitivity of 91.9%, 93.0%, and 90.7%, respectively.

Torkzaban *et al*. (2015) explored an innovative approach for phylogeographic analysis and the genealogical study of olive populations by employing machine learning techniques to classify microsatellite markers (Torkzaban *et al*. 2015). In the feature selection phase of the weighting algorithm, they used the Gini index, correlation, information acquisition, SVM, and genetic algorithm (selection optimizer). The weighting algorithm was used in the feature selection phase, and the attributes with a weight equal to or greater than 0.5 were selected by each algorithm and stored as a new feature-weighted dataset. Then, the data were organized. At the final stage, i.e., evaluating classification methods, prediction methods of tree induction and Naive Bayes were employed, and an accuracy of 84.30% was obtained. In our research, the SVM algorithm provided better results than other techniques in predicting genes encoding AMPs at the appropriate genomic level. SVM is a binary classification algorithm that distinguishes between two classes by projecting feature data points onto a multidimensional space and establishing a hyperplane to divide them. This hyperplane is strategically placed to maximize the margin between the two classes, ensuring optimal classification when applied to a test set. In bioinformatics, the SVM weighting algorithm is one of the most commonly used supervised learning techniques, thanks to its robust statistical foundation (Meher *et al*. 2017). Over the past decade, there has been a significant surge in the development of machine learning models aimed at predicting antimicrobial peptides (AMPs).

Among the most frequently used classifiers for AMP prediction are support vector machines (SVM) and random forests (RF) (Lata *et al*. 2010; Joseph *et al*. 2012; Porto *et al*. 2012; Thakur *et al*. 2012; Qureshi *et al*. 2015; Meher *et al*. 2017; Bhadra *et al*. 2018; Jhong *et al*. 2019; Burdukiewicz *et al*. 2020; Kavousi *et al*. 2020; Santos-Junior *et al*. 2020; Sharma *et al*. 2021) . The output datasets were confirmed by performing the weighting algorithms in two stages for training and optimizing the third data (forecasting data), consisting of 10 and 22 recessions, respectively. The results and observations from the output of the SVM machine learning algorithm show an accuracy rate of 99.63%. SVM, RF, and artificial neural networks have been widely used in previous studies to identify AMPs (Shah *et al*. 2017). In this research, the results obtained from the iAMPpred online server were analyzed to confirm the prediction accuracy of genes encoding AMPs in our work (Meher *et al*. 2017). The performance of iAMPpred was evaluated further to predict antimicrobial peptides (AMPs) specific to four distinct source organisms: humans, sheep,

bovine, and Gallus gallus. The tool demonstrated high accuracy in identifying peptide sequences with antibacterial and antiviral properties.

The proposed approach is considered to build upon and enhance the tools and techniques previously developed for AMP prediction. Accordingly, the peptide-encoding genes predicted by the algorithm were converted into protein sequences, and their antimicrobial properties were investigated in silico using the mentioned web server. This analysis revealed that the weighting algorithms used in this study were able to predict the genes encoding antimicrobial peptides successfully. We investigated these polypeptides (i.e., ten genes that coded antimicrobial peptides in the genomes of three animal species and humans as AMPs) for which there is no laboratory confirmation in silico. The results showed that these polypeptides can be AMPs suitable for the purpose of this research. Aside from its antimicrobial properties, AMPs have anticancer properties and stimulate the immune system.

As previous studies have focused on the anticancer aspect of these peptides, they may have multifunctional and antimicrobial properties. Polypeptides that are predicted to have antimicrobial properties, in addition to their well-known antimicrobial properties, could contribute to preventing cancer and regulating the immune system (Duarte-Mata and Salinas-Carmona, 2023). The heightened sensitivity of tumor cells to membrane-active cationic antimicrobial peptides (AMPs) stems from the elevated presence of anionic phosphatidylserine molecules on their membranes compared to normal cells. This characteristic has paved the way for utilizing AMPs as effective antitumor agents (Utsugi *et al*. 1991; Liu *et al*. 2015). We predicted ten genes that can be Amp (Table 6). In past bioinformatics studies, ten genes that we predicted as AMPs were introduced as prognostic genes in cancer. POLR2K gene acts as the bladder cancer prognosis. The surface of the cancer cell membrane offers enhanced potential for AMPs, largely attributed to its expanded surface area resulting from the increased presence of microvilli (Munteanu *et al*. 2009). Cancer cells exhibit reduced levels of cholesterol-based anchors, making them more vulnerable to antimicrobial peptides (AMPs) (Repana *et al*. 2019). Ha *et al*. (2021) reported that GNAi2/gip2 transcriptome could stimulate ovarian cancer growth (Ha *et al*. 2021). They found that advanced ovary cancer could be diagnosed and treated considering gene expression at the last stage as well as hub and bottleneck nodes such as UQCRFS. COX7A1 is the subunit of cytochrome c oxidase. Feng *et al*. (2022) explored the relationship between COX7A1 and ferroptosis, a recently identified form of cell death driven by iron-dependent lipid peroxidation, across various human non-small-cell lung carcinoma (NSCLC) cell lines. Researchers discovered that COA enhances the sensitivity of NSCLC cells to ferroptosis triggered by cysteine deprivation. This effect is achieved by boosting the activity of the tricarboxylic acid (TCA) cycle and complex IV within the mitochondrial electron transport chain (ETC). Previous studies have revealed AMPs could inhibit the growth of lung carcinoma cell lines (Kunda, 2020).

Moreover, cancer cells might exhibit a stronger negative charge than normal cells, potentially attributed to the excessive expression of glycoproteins or glycosaminoglycans on their membranes. This increased surface charge could enhance the binding affinity of AMPs (Peduzzi *et al*. 1996; Dos Santos *et al*. 2017). Small nuclear ribonucleoprotein polypeptides B and B1 (SNRPB) are essential elements of the spliceosome, significantly contributing to the pre-mRNA splicing process (Gray *et al*. 1999). Dysregulation of SNRPB disrupts pre-mRNA splicing, leading to the production of unusual mRNA variants. The proteins derived from these atypical mRNA variants could have a substantial impact on tumorigenesis (Correa *et al*. 2016; Peng *et al*. 2020; Zhan *et al*. 2021). The role of SNR in tumor initiation and progression has been identified in various cancers, including NSCLC (Liu *et al*. 2019; Liu *et al*. 2021). The ubiquitin-like protein UBD, which plays a major role in cell proliferation and sister chromatid cohesion through associating with spliceosomes, participates in pre-mRNA splicing by maintaining spliceosome integrity. UBD significantly contributes to the degradation of various proteins, cell differentiation, cell cycle regulation, apoptosis, message transmission, DNA repair, stress response, and immune response, all of which are not performed by protein degradation. UBD expression level influences tumor development (Lukasiak *et al*. 2008; Izadi *et al*. 2014). The mitochondrial ribosomal protein (MRP) family is crucial for mitochondrial energy metabolism, a core aspect influencing the development of breast cancer. Among these proteins, MRPL33 is essential for maintaining mitochondrial function and plays a significant role in promoting tumor progression (Wallace, 2012; Gatza *et al*. 2014; Li *et al*. 2019). Therefore, it is possible that these polypeptides in our study can be AMPs, although experimental tests are needed for 100% confirmation. The present research investigated AMP prediction using simultaneous machine learning algorithms in humans and three animal species at the protein sequence and DNA sequence levels. Previous studies have examined prediction exclusively at the amino acid sequence level within a single species (Torrent *et al*. 2011; Bhadra *et al*. 2018).

CONCLUSION

In this paper, Using a current dataset, a machine learning model was developed to predict the AMP of antimicrobial peptides with excellent performance. Feature selection with cross-validation on a set of physical and chemical features at the protein sequence and DNA sequence level was used to identify genes encoding AMP antimicrobial peptides in the genomes of three strategic animals and humans. The most groundbreaking aspect of this approach was determined to be genome-level motifs in the intron region, 1 kb upstream, and 1 kb downstream of the gene. The feature of genome-level motifs was used for the first time in this study to predict AMP. It was revealed that genome-level motifs play an important role in peptide function thus, It should be taken into account when training new models. The outcomes and observations from the output of the machine learning algorithm of this study showed that the optimal Naive Bayes model processed with the SVM training data set with an accuracy of 99.63% was recognized as the best model and performed well compared to other models. The explanation is that this model predicted the gene that encodes peptides with antimicrobial properties with an accuracy of 96.63% (Table 6).

ACKNOWLEDGEMENT

We would like to express our profound gratitude to Ferdowsi University of Mashhad for support of this project.

REFERENCES

Alelyani S. (2021). Stable bagging feature selection on medical data. *J. Big Data*. ***8*(1),** 1-18.

Altman N. and Krzywinski M. (2017). Points of significance: clustering. *Nat. Methods*. **14(6),** 545-547.

Andersson E., Rydengård V., Sonesson A., Mörgelin M., Björck L. and Schmidtchen A. (2004). Antimicrobial activities of heparin-binding peptides. *European J. Biochem*. **271(6),** 1219-1226.

Attique M., Farooq M.S., Khelifi A. and Abid A. (2020). Prediction of therapeutic peptides using machine learning: computational models, datasets, and feature encodings. *IEEE Access*. **8,** 148570-148594.

Barahimipour R., Strenkert D., Neupert J., Schroda M., Merchant S.S. and Bock R. (2015). Dissecting the contributions of GC content and codon usage to gene expression in the model alga Chlamydomonas reinhardtii. *Plant J*. **84(4),** 704-717.

Belgiu M. and Drăguţ L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm Remote Sens.* **114,** 24-31.

Bhadra P., Yan J., Li J., Fong S. and Siu S.W. (2018). AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.* **8(1),** 1697-1703.

Boris V., Maya G., Grigol M., Gabrielian A., Rosenthal A. and Darrell E.H. (2022). Comparative analysis of machine learning algorithms on the microbial strain-specific AMP prediction. *J. Brief Bioinform*. **23(4),** 233-241.

Breiman L. (2001). Random forests. *Machine Learn*. **45,** 5-32.

Buchan D.W. and Jones D.T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47(1),** 402-407.

Burdukiewicz M., Sidorczuk K., Rafacz D., Pietluch F., Chilimoniuk J., Rödiger S. and Gagat P. (2020). Proteomic screening for prediction and design of antimicrobial peptides with AmpGram. *Int. J Mol. Sci.* **21(12),** 4310-4321.

Cabras T., Patamia M., Melino S., Inzitari R., Messana I., Castagnola M. and Petruzzelli R. (2007). Pro-oxidant activity of histatin 5 related Cu (II)-model peptide probed by mass spectrometry. *Biochem. Biophys. Res. Commun*. **358(1),** 277-284.

Cai B. and Jiang X. (2016a). Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences. *BMC Bioinformatics*. **17,** 1-12.

Cai B. and Jiang X. (2016b). Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences. *BMC Bioinformatics*. **17(1),** 1-12.

Cao D.S., Xu Q.S. and Liang Y.Z. (2013). propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics.* **29(7),** 960-962.

Cardoso P., Glossop H., Meikle T.G., Aburto-Medina A., Conn C. E., Sarojini V. and Valery C. (2021). Molecular engineering of antimicrobial peptides: Microbial targets, peptide motifs and translation opportunities. *Biophys. Rev*. **13,** 35-69.

Chapelle O., Haffner P. and Vapnik V.N. (1999). Support vector machines for histogram-based image classification. *IEEE Trans Neural Network. Learn. Syst.* **10(5),** 1055-1064.

Cheema U., Younas M., Sultan J., Iqbal A., Tariq M. and Waheed A. (2011). Antimicrobial peptides: An alternative of antibiotics in ruminants. *Adv. Agric. Biotechnol*. **2,** 15-21.

Cools T.L., Struyfs C., Cammue B.P. and Thevissen K. (2017). Antifungal plant defensins: Increased insight in their mode of action as a basis for their use to combat fungal infections. *Future Microbiol.* **12(5),** 441-454.

Correa B.R., de Araujo P.R., Qiao M., Burns S.C., Chen C., Schlegel R., Agarwal S., Galante P.A. and Penalva L.O. (2016). Functional genomics analyses of RNA-binding proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma. *Genome Biol.* **17,** 1-16.

Cortes C. and Vapnik V. (1995). Support-vector networks. *Machine Learn*. **20,** 273-297.

Cowley M., de Burca A., McCole R.B., Chahal M., Saadat G., Oakey R.J. and Schulz R. (2011). Short interspersed element (SINE) depletion and long interspersed element (LINE) abundance are not features universally required for imprinting. *PLoS One*. **6(4),** e18953.

De Groot N.S., Castillo V., Graña-Montes R. and Ventura S. (2012). AGGRESCAN: method, application, and perspectives for drug design. *Methods Mol. Biol*. **819,** 199-220.

D'haeseleer P. (2006). What are DNA sequence motifs? *Nat. Biotechnol.* **24(4),** 423-425.

De la Fuente-Núñez C., Silva O.N., Lu T.K. and Franco O.L. (2017). Antimicrobial peptides: Role in human disease and potential as immunotherapies. *Pharmacol. Ther*. **178,** 132-140.

Dos Santos C., Hamadat S., Le Saux K., Newton C., Mazouni M., Zargarian L., Miro-Padovani M., Zadigue P., Delbé J. and Hamma-Kourbali Y. (2017). Studies of the antitumor mechanism of action of dermaseptin B2, a multifunctional cationic antimicrobial peptide, reveal a partial implication of cell surface glycosaminoglycans. *PLoS One*. **12(8),** e0182926.

Duarte-Mata D.I. and Salinas-Carmona M.C. (2023). Antimicrobial peptides immune modulation role in intracellular bacterial infection. *Front Immunol*. **14,** 1119574-1119583.

Erdem Büyükkiraz M. and Kesmen Z. (2022). Antimicrobial peptides (AMPs): A promising class of antimicrobial compounds. *J. Appl. Microbiol.* **132(3),** 1573-1596.

Fan L., Sun J., Zhou M., Zhou J., Lao X., Zheng H. and Xu H. (2016). DRAMP: A comprehensive data repository of antimicrobial peptides. *Sci. Rep.* **6(1),** 24482-24491.

Feng Y., Xu J., Shi M., Liu R., Zhao L., Chen X., Li M., Zhao Y., Wenjing J.C. and Liu P. (2022). COX7A1 enhances the sensitivity of human NSCLC cells to cystine deprivation-induced ferroptosis via regulating mitochondrial metabolism. *Cell Death Dis*. **13(11),** 988-997.

Ganz T. (2007). Biosynthesis of defensins and other antimicrobial peptides. *Ciba Found. Symp*. **186,** 62-76.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S.E., Wilkins M.R., Appel R.D. and Bairoch A. (2005). Protein identification and analysis tools on the ExPASy server. *Methods Mol. Biol*. **571,** 607-715.

Gatza M.L., Silva G.O., Parker J.S., Fan C. and Perou C.M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet*. **46(10),** 1051-1059.

Gray T.A., Smithwick M.J., Schaldach M.A., Martone D.L., Marshall Graves J.A., McCarrey J.R. and Nicholls R.D. (1999). Concerted regulation and molecular evolution of the duplicated SNRPB'/B and SNRPN Ioci. *Nucleic Acids Res*. **27(23),** 4577-4584.

Gudmundsson G.H., Magnusson K.P., Chowdhary B.P., Johansson M., Andersson L. and Boman H.G. (1995). Structure of the gene for porcine peptide antibiotic PR-39, a cathelin gene family member: comparative mapping of the locus for the human peptide antibiotic FALL-39. *Proc. Natl. Acad Sci.* **92(15),** 7085-7089.

Ha J.H., Jayaraman M., Yan M., Dhanasekaran P., Isidoro C., Song Y.S. and Dhanasekaran D.N. (2021). GNAi2/gip2-regulated transcriptome and its therapeutic significance in ovarian cancer. *Biomolecules*. **11(8),** 1211-1224.

Hancock R.E., Haney E.F. and Gill E.E. (2016). The immunology of host defence peptides: beyond antimicrobial activity. *Nat. Rev. Immunol.* **16(5),** 321-334.

Herbinière J., Braquart-Varnier C., Grève P., Strub J.M., Frère J., Van Dorsselaer A. and Martin G. (2005). Armadillidin: a novel glycine-rich antibacterial peptide directed against gram-positive bacteria in the woodlouse Armadillidium vulgare (Terrestrial Isopod, Crustacean). *Dev. Comp. Immunol*. **29(6),** 489-499.

Hosseinzadeh F., Ebrahimi M., Goliaei B. and Shamabadi N. (2012). Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. *PLoS One*. **7(7),** e40017.

Ivanova M.I., Sawaya M.R., Gingery M., Attinger A. and Eisenberg D. (2004). An amyloid-forming segment of β2-microglobulin suggests a molecular model for the fibril. *Proc. Natl. Acad Sci. USA.* **101(29),** 10584-10589.

Izadi A., Moslemi E., Poorhosseini S.M., Yassaee V.R., Kheiri H.R. and Elikai H.R. (2014). UBD identify in paraffin tissues in patients with colorectal cancer. *J. Isfahan Med. Sch*. **32(291),** 1-10.

Jhong J.H., Chi Y.H., Li W.C., Lin T.H., Huang K.Y. and Lee T.Y. (2019). dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Res*. **47(1),** 285-297.

Joseph S., Karnik S., Nilawe P., Jayaraman V.K. and Idicula-Thomas S. (2012). ClassAMP: A prediction tool for classification of antimicrobial peptides. *IEEE/ACM Trans Comput Biol Bioinform*. **9(5),** 1535-1538.

Karami K., Zerehdaran S., Javadmanesh A., Shariati M.M. and Fallahi H. (2019). Characterization of bovine (*Bos taurus*) imprinted genes from genomic to amino acid attributes by data mining approaches. *PLoS One*. **14(6),** e0217813.

Katoch S., Chauhan S.S. and Kumar V. (2021). A review on genetic algorithm: past, present, and future. *Multimed Tools Appl.* **80,** 8091-8126.

Kavousi K., Bagheri M., Behrouzi S., Vafadar S., Atanaki F.F., Lotfabadi B.T., Ariaeenejad S., Shockravi A. and Moosavi-Movahedi A.A. (2020). IAMPE: NMR-assisted computational prediction of antimicrobial peptides. *J. Chem. Inf. Model.* **60(10),** 4691-4701.

Kelley D.R., Snoek J. and Rinn J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26(7),** 990-999.

Khabbaz H., Karimi-Jafari M.H., Saboury A.A. and BabaAli B. (2021). Prediction of antimicrobial peptides toxicity based on their physico-chemical properties using machine learning techniques. *BMC Bioinform*. **22(1),** 1-11.

Kondori N., Baltzer L., Dolphin G. and Mattsby-Baltzer I. (2011). Fungicidal activity of human lactoferrin-derived peptides based on the antimicrobial αβ region. *Int. J. Antimicrob. Agents.* **37(1),** 51-57.

Kuiken T. (2023). Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight, U.S. Congressional Research Service, November 22, 2023, https://crsreports.congress.gov/product/pdf/R/R47849.

Kunda N.K. (2020). Antimicrobial peptides as novel therapeutics for non-small cell lung cancer. *Drug Discov. Today Technol.* **25(1),** 238-247.

Lai Y. and Gallo R.L. (2009). AMPed up immunity: How antimicrobial peptides have multiple roles in immune defense. *Trends Immunol*. **30(3),** 131-141.

Lata S., Mishra N.K. and Raghava G.P. (2010). AntiBP2: improved version of antibacterial peptide prediction. *BMC*

Bioinform. **11,** 1-7.

Lata S., Sharma B. and Raghava G.P. (2007). Analysis and prediction of antibacterial peptides. *BMC Bioinform*. **8,** 1-10.

Lay F. and Anderson M. (2005). Defensins-components of the innate immune system in plants. *Curr. Protein Pept. Sci*. **6(1),** 85-101.

Lee E.Y., Fulan B.M., Wong G.C. and Ferguson A.L. (2016). Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proc. Natl. Acad Sci. USA*. **113(48),** 13588-13593.

Leptihn S., Har J.Y., Wohland T. and Ding J.L. (2010). Correlation of charge, hydrophobicity, and structure with antimicrobial activity of S1 and MIRIAM peptides. *Biochemistry*. **49(43),** 9161-9170.

Li J., Feng D., Gao C., Zhang Y., Xu J., Wu M. and Zhan X. (2019). Isoforms S and L of MRPL33 from alternative splicing have isoform-specific roles in the chemoresponse to epirubicin in gastric cancer cells via the PI3K/AKT signaling pathway. *Int. J. Oncol*. **54(5),** 1591-1600.

Liu N., Chen A., Feng N., Liu X. and Zhang L. (2021). SNRPB is a mediator for cellular response to cisplatin in non-small-cell lung cancer. *Medical Oncol.* **38,** 1-9.

Liu N., Wu Z., Chen A., Wang Y., Cai D., Zheng J., Liu Y. and Zhang L. (2019). SNRPB promotes the tumorigenic potential of NSCLC in part by regulating RAB26. *Cell Death Dis*. **10(9),** 667-678.

Liu S., Fan L., Sun J., Lao X. and Zheng H. (2017). Computational resources and tools for antimicrobial peptides. *J. Pept. Sci.* **23(1),** 4-12.

Liu X., Li Y., Li Z., Lan X., Leung P. H.M., Li J., Yang M., Ko F. and Qin L. (2015). Mechanism of anticancer effects of antimicrobial peptides. *Int. J. Med. Inform.* **8(1),** 25-36.

Lukasiak S., Schiller C., Oehlschläger P., Schmidtke G., Krause P., Legler D.F., Autschbach F., Schirmacher P., Breuhahn K. and Groettrup M. (2008). Proinflammatory cytokines cause FAT10 upregulation in cancers of liver and colon. *Oncogene*. **27(46),** 6068-6074.

Marina P.S., Ekaterina A.I., Tatyana V.K. and Tatyana I.O. (2023). The γ-Core motif peptides of plant AMPs as novel antimicrobials for medicine and agriculture. *Int. J. Mol. Sci*. **24(1),** 483-491.

Martin E., Ganz T. and Lehrer R.I. (1995). Defensins and other endogenous peptide antibiotics of vertebrates. *J. Leukoc. Biol.* **58(2),** 128-136.

Meher P.K., Sahu T.K., Saini V. and Rao A.R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **7(1),** 42362-42372.

Min C., Ohta K., Kajiya M., Zhu T., Sharma K., Shin J., Mawardi H., Howait M., Hirschfeld J. and Bahammam L. (2012). The antimicrobial activity of the appetite peptide hormone ghrelin. *Peptides*. **36(2),** 151-156.

Mueller A., Candrian G., Kropotov J.D., Ponomarev V.A. and Baschera G.M. (2010). Classification of ADHD patients on the basis of independent ERP components using a machine learning system. *Nonlinear Biomed Phys*. **4,** 1-12.

Munteanu C.R., Magalhães A.L., Uriarte E. and González-Díaz H. (2009). Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.* **257(2),** 303-311.

Pane K., Durante L., Crescenzi O., Cafaro V., Pizzo E., Varcamonti M., Zanfardino A., Izzo V., Di Donato A. and Notomista E. (2017). Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: Application to the detection of "cryptic" antimicrobial peptides. *J. Theor. Biol.* **419,** 254-265.

Papagianni M. (2003). Ribosomally synthesized peptides with antimicrobial properties: biosynthesis, structure, function, and applications. *Biotechnol. Adv.* **21(6),** 465-499.

Peduzzi P., Concato J., Kemper E., Holford T.R. and Feinstein A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol*. **49(12),** 1373-1379.

Peng N., Li J., He J., Shi X., Huang H., Mo Y., Ye H., Wu G., Wu F. and Xiang B. (2020). c□Myc□mediated SNRPB upregulation functions as an oncogene in hepatocellular carcinoma. *Cell Biol. Int.* **44(5),** 1103-1111.

Peterson A.T., Papeş M. and Soberón J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecol. Model.* **213(1),** 63-72.

Piryonesi S.M. and El-Diraby T.E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. *J. Infrastruct. Syst.* **26(1),** 1-10.

Porto W.F., Pires Á.S. and Franco O.L. (2012). CS-AMPPred: An updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. *PLoS One*. **7(12),** e51444.

Qureshi A., Tandon H. and Kumar M. (2015). AVP- IC50Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50). *Peptide Sci*. **104(6),** 753-763.

Repana D., Nulsen J., Dressler L., Bortolomeazzi M., Venkata S.K., Tourna A., Yakovleva A., Palmieri T. and Ciccarelli F.D. (2019). The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.* **20,** 1-12.

Richardson J.S. (1994). Introduction: protein motifs. *FASEB J.* **8(15),** 1237-1239.

Richardson S.R., Doucet A.J., Kopera H.C., Moldovan J.B., Garcia-Perez J.L. and Moran J.V. (2015). The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Mobile DNA III*. ASM Press., Washington, DC, USA.

Rončević T., Puizina J. and Tossi A. (2019). Antimicrobial peptides as anti-infective agents in pre-post-antibiotic era? *Int. J. Mol. Sci.* **20(22),** 5713-5721.

Sagaram U.S., Kaur J. and Shah D. (2012). Antifungal plant defensins: structure -activity relationships, mode of action, and biotech applications. Pp. 317-336 in Small Wonders: Peptides for Disease Control. K. Rajasekaran, Ed., Washinton, DC, USA.

Santos-Junior C.D., Pan S., Zhao X.M. and Coelho L.P. (2020). Macrel: Antimicrobial peptide screening in genomes and

metagenomes. *Peer J*. **8,** e10555.
Schauber J. and Gallo R.L. (2007). Expanding the roles of antimicrobial peptides in skin: Alarming and arming keratinocytes. *J. Invest. Dermatol.* **127(3),** 510-512.
Schlesinger D. and Elsässer S.J. (2022). Revisiting sORFs: Overcoming challenges to identify and characterize functional microproteins. *FEBS J*. **289(1),** 53-74.
Selsted M.E. and Ouellette A.J. (1995). Defensins in granules of phagocytic and non-phagocytic cells. *Trends Cell Biol.* **5(3),** 114-119.
Shah Y., Sehgal D. and Valadi J.K. (2017). Recent trends in antimicrobial peptide prediction using machine learning techniques. *Bioinformation*. **13(12),** 415-425.
Sharma R., Shrivastava S., Kumar Singh S., Kumar A., Saxena S. and Kumar Singh R. (2021). AniAMPpred: Artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom. *Briefings in Bioinform*. **22(6),** 242-256.
Söylemez Ü.G., Yousef M. and Bakir-Gungor B. (2023). Prediction of antimicrobial peptides using deep neural networks. *Bioinformation.* **2023,** 188-194.
Tay W.M., Hanafyv A.I., Angerhoferv A. and Ming L.J. (2009). A plausible role of salivary copper in antimicrobial activity of histatin-5—metal binding and oxidative activity of its copper complex. *Bioorg. Med. Chem. Lett.* **19(23),** 6709-6712.
Thakur N., Qureshi A. and Kumar M. (2012). AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res.* **40(1),** 199-204.
Thomas S., Karnik S., Barai R.S., Jayaraman V.K. and Thomas S.I. (2010). CAMP: A useful resource for research on antimicrobial peptides. *J. Nucleic Acids Res*. **38(1),** 774-780.
Torkzaban B., Kayvanjoo A.H., Ardalan A., Mousavi S., Mariotti R., Baldoni L., Ebrahimie E., Ebrahimi M. and Hosseini M. (2015). Machine learning based classification of microsatellite variation: an effective approach for phylogeographic characterization of olive populations. *PLoS One*. **10(11),** e0143465.
Torrent M., Andreu D., Nogués V.M. and Boix E. (2011). Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS One*. **6(2),** e16968.
Trivedi N. K., Tiwari R. G., Anand A., Gautam V., Witarsyah D. and Misra A. (2022). Application of machine learning for diagnosis of liver cancer 2022 Pp. 1-5 in Int. Conf. Adv. Data Sci., E-Learn. Inform. Syst., Bandung, Indonesia.
Utsugi T., Schroit A.J., Connor J., Bucana C.D. and Fidler I.J. (1991). Elevated expression of phosphatidylserine in the outer membrane leaflet of human tumor cells and recognition by activated human blood monocytes. *Cancer Res.* **51(11),** 3062-3066.
Vassetzky N.S. and Kramerov D.A. (2013). SINEBase: A database and tool for SINE analysis. *Nucleic Acids Res.* **41(1),** 83-89.
Ventura S., Zurdo J., Narayanan S., Parreño M., Mangues R., Reif B., Chiti F., Giannoni E., Dobson C.M. and Aviles F.X. (2004). Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl. Acad. Sci. USA.* **101(19),** 7258-7263.
Vishnepolsky B., Grigolava M., Managadze G., Gabrielian A., Rosenthal A., Hurt D.E., Tartakovsky M. and Pirtskhalava M. (2022). Comparative analysis of machine learning algorithms on the microbial strain-specific AMP prediction. *Brief. Bioinform.* **23(4),** 233-242.
Vishnepolsky B. and Pirtskhalava M. (2014). Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes. *J. Chem. Inf. Model.* **54(5),** 1512-1523.
Waghu F.H., Barai R.S., Gurung P. and Thomas S.I. (2016). CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res*. **44(1),** 1094-1097.
Wallace D.C. (2012). Mitochondria and cancer. *Nat. Rev. Cancer.* **12(10),** 685-698.
Wang G. (2022). Unifying the classification of antimicrobial peptides in the antimicrobial peptide database. *Methods Enzymol*. **663,** 1-18.
Xie C., Prahl A., Ericksen B., Wu Z., Zeng P., Li X., Lu W.Y., Lubkowski J. and Lu W. (2005). Reconstruction of the conserved β-bulge in mammalian defensins using D-amino acids. *J. Biol. Chem.* **280(38),** 32921-32929.
Yamagata N., Shyr Y., Yanagisawa K., Edgerton M., Dang T.P., Gonzalez A., Nadaf S., Larsen P., Roberts J.R. and Nesbitt J.C. (2003). A training-testing approach to the molecular classification of resected non-small cell lung cancer. *Clin. Cancer Res.* **9(13),** 4695-4704.
Yeaman M.R. and Yount N.Y. (2003). Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev*. **55(1),** 27-55.
Yount N.Y. and Yeaman M.R. (2004). Multidimensional signatures in antimicrobial peptides. *Proc. Natl. Acad. Sci. USA.* **101(19),** 7363-7368.
Zasloff M. (2002). Antimicrobial peptides of multicellular organisms. *Nature*. **415(6870),** 389-395.
Zhan Y.T., Li L., Zeng T.T., Zhou N.N., Guan X.Y. and Li Y. (2021). SNRPB-mediated RNA splicing drives tumor cell proliferation and stemness in hepatocellular carcinoma. *Aging.* **13(1),** 537.
Zhao C., Ganz T. and Lehrer R.I. (1995). Structures of genes for two cathelin- associated antimicrobial peptides: prophenin- 2 and PR- 39. *FEBS Lett.* **376(3),** 130-134.
Zhao X., Wu H., Lu H., Li G. and Huang Q. (2013). LAMP: A database linking antimicrobial peptides. *PLoS One*. **8(6),** e66557.